Development and validation of multivariable prediction models of remission, recovery, and quality of life outcomes in people with first episode psychosis: a machine learning approach

Samuel P Leighton, Rachel Upthegrove, Rajeev Krishnadas, Michael E Benros, Matthew R Broome, Georgios V Gkoutos, Peter F Liddle, Swaran P Singh, Linda Everard, Peter B Jones, David Fowler, Vimal Sharma, Nicholas Freemantle, Rune H B Christensen, Nikolai Albert, Merete Nordentoft, Matthias Schwannauer, Jonathan Cavanagh, Andrew I Gumley, Max Birchwood, Pavan K Mallikarjun**

Summary

Background Outcomes for people with first-episode psychosis are highly heterogeneous. Few reliable validated methods are available to predict the outcome for individual patients in the first clinical contact. In this study, we aimed to build multivariable prediction models of 1-year remission and recovery outcomes using baseline clinical variables in people with first-episode psychosis.

Methods In this machine learning approach, we applied supervised machine learning, using regularised regression and nested leave-one-site-out cross-validation, to baseline clinical data from the English Evaluating the Development and Impact of Early Intervention Services (EDEN) study (n=1027), to develop and internally validate prediction models at 1-year follow-up. We assessed four binary outcomes that were recorded at 1 year: symptom remission, social recovery, vocational recovery, and quality of life (QoL). We externally validated the prediction models by selecting from the top predictor variables identified in the internal validation models the variables shared with the external validation datasets comprised of two Scottish longitudinal cohort studies (n=162) and the OPUS trial, a randomised controlled trial of specialised assertive intervention versus standard treatment (n=578).

Findings The performance of prediction models was robust for the four 1-year outcomes of symptom remission (area under the receiver operating characteristic curve [AUC] 0·703, 95% CI 0·664–0·742), social recovery (0·731, 0·697–0·765), vocational recovery (0·736, 0·702–0·771), and QoL (0·704, 0·667–0·742; p<0·0001 for all outcomes), on internal validation. We externally validated the outcomes of symptom remission (AUC 0·680, 95% CI 0·587–0·773), vocational recovery (0·867, 0·805–0·930), and QoL (0·679, 0·522–0·836) in the Scottish datasets, and symptom remission (0·616, 0·553–0·679), social recovery (0·573, 0·504–0·643), vocational recovery (0·660, 0·610–0·710), and QoL (0·556, 0·481–0·631) in the OPUS dataset.

Interpretation In our machine learning analysis, we showed that prediction models can reliably and prospectively identify poor remission and recovery outcomes at 1 year for patients with first-episode psychosis using baseline clinical variables at first clinical contact.

Funding Lundbeck Foundation.

Introduction

Psychosis is an illness with an early first onset, occurring usually in young people and with an incidence of 31 per 100 000 person-years.[1] Patients with first-episode psychosis have heterogeneity of outcomes, with a 58% prevalence of remission and 38% of recovery.[2] The identification of individual-patient outcomes at initial clinical contact might help to personalise treatment and lead to improved use of resources for those most in need or likely to respond to treatment.[3] However, few validated tools are available for the accurate early identification of patients with good or poor outcomes.

Previous observational studies have identified predictors of outcomes at the group level, including sociodemographic factors, clinical and treatment response variables, comorbidity, and functional and cognitive deficits,[2,4,5] with inconsistent reliability.[6] More clarity is needed on how to apply group-level factors to an individual level of prediction. An approach that can be applied to stratify the individualised risk of a poor outcome at the initial clinical contact is required. One solution is the use of machine learning, in which algorithms can sift through a large array of predictor variables and detect complex high dimensional interactions that can reliably predict individual-patient outcomes.[7]

Lancet Digital Health 2019

Published **Online** September 12, 2019 http://dx.doi.org/10.1016/S2589-7500(19)30121-9

See **Online/Comment** http://dx.doi.org/10.1016/S2589-7500(19)30122-0

*Joint senior authors

Institute of Health and Wellbeing (S P Leighton MBChB, Prof J Cavanagh MD, Prof A I Gumley PhD) **and Institute of Neuroscience and Psychology** (R Krishnadas PhD), **University of Glasgow, Glasgow, UK; Institute for Mental Health** (Prof R Upthegrove PhD, Prof M R Broome PhD, P K Mallikarjun PhD), **Institute of Cancer and Genomics** (Prof G V Gkoutos PhD), **and Institute of Translational Medicine** (Prof G V Gkoutos), **University of Birmingham, Birmingham, UK; Copenhagen Research Center for Mental Health, Mental Health Centre Copenhagen, Copenhagen University Hospital, Copenhagen, Denmark** (M E Benros PhD, R H B Christensen PhD, N Albert PhD, Prof M Nordentoft PhD); **Health Data Research UK Midlands, UK** (Prof G V Gkoutos); **University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK** (Prof G V Gkoutos); **Institute of Mental Health, University of Nottingham, Nottingham, UK** (Prof P F Liddle PhD); **Mental Health and Wellbeing, Warwick Medical School, University of Warwick, Coventry, UK** (Prof S P Singh MD, Prof M Birchwood PhD); **The Barberry, Birmingham, UK** (L Everard BSc); **Wolfson College, University of**

Cambridge, Cambridge, UK (Prof P B Jones MD); **School of Psychology, University of Sussex, Brighton, UK** (Prof F Fowler MSc); **Department of Health and Social Care, University of Chester, Chester, UK** (Prof V Sharma PhD); **Comprehensive Trials Unit, University College London, London, UK** (Prof N Freemantle PhD); **Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark** (Prof M Nordentoft); **and School of Health in Social Science, Clinical Psychology, University of Edinburgh, Edinburgh, UK** (Prof M Schwannauer PhD)

Correspondence to:
Dr Pavan Mallikarjun, Institute for Mental Health, University of Birmingham, Birmingham B15 2SA, UK
p.mallikarjun@bham.ac.uk

Research in context

Evidence before this study

In patients with first-episode psychosis, prediction of remission and recovery outcomes is an important goal during initial clinical contact. These patients have heterogeneous outcomes, even with standardised interventions. Targeting extended or more intensive treatment to patients with poorer prognosis might lead to better outcomes. Previous studies have identified several group-level predictors, including poor premorbid adjustment, history of developmental disorder, symptom severity at baseline, and duration of untreated psychosis, as predictors of poor clinical, functional, and cognitive outcomes. Such group-level differences are not always replicated at the individual level, and how to combine the group-level factors for individualised prediction is unclear. We searched PubMed from inception to March 12, 2019, using the terms "psychosis" or "first episode psychosis" or "schizophrenia" AND "prediction" AND "outcome" in any field, with no language restrictions. We retrieved 470 articles, of which, after excluding articles not related to multivariable prediction of outcomes based on baseline clinical variables, we identified two articles that have published models for outcome prediction in psychosis using baseline variables. One study had developed an internally cross-validated model for prediction of functional outcomes in a large sample, but this model was not externally validated on an independent sample. The other study developed remission and recovery prediction models on a small sample from a longitudinal cohort study, and externally validated the models on patients from a different cohort study. Additionally, examples exist of outcome prediction models that have been internally cross-validated and externally validated for depression.

Added value of this study

To our knowledge, our study provides the first reliable evidence for the usefulness of machine learning to develop outcome prediction models, using baseline variables at first clinical contact, in a large sample of patients with first-episode psychosis. The models use baseline clinical and demographic data, rather than neuroimaging or other biomarkers, and, as such, are more accessible in a clinical setting for potential future applications. Our results were validated by the methods that we used, including internal–external validation of the outcome prediction models developed on data from a large multicentre cohort study and external validation on a small cohort study of patients with first-episode psychosis and a large randomised controlled trial of patients with first-episode psychosis. Our study attempted to develop outcome prediction models for multiple outcomes (clinical, recovery, and quality of life), although a single model might be useful to predict multiple outcomes, albeit with reduced accuracy.

Implications of all the available evidence

Our study, and the two previous studies, showed that machine learning techniques applied to baseline clinical and demographic data can aid in the prediction of remission and recovery outcomes for patients with first-episode psychosis at first clinical contact. This approach can be extended to include other sources of data (neuroimaging data, immune biomarkers, and so on), which might enhance model performance. The next step before implementation into routine clinical practice would be to investigate the usefulness of the prediction models in prospective controlled trials.

Two models developed for outcome prediction in psychosis using baseline variables have been published.[8,9] Koutsouleris and colleagues[8] used machine learning to predict 4-week and 52-week functional outcomes in patients with first-episode psychosis to a 75·0% (for 4 weeks) and 73·8% (for 52 weeks) test-fold balanced accuracy (ie, average accuracy across the ten folds) on repeated nested internal cross-validation, with use of data from a randomised control study (n=334); however, this model was not externally validated. Leighton and colleagues[9] developed 1-year remission and vocational recovery prediction models on 83 patients with first-episode psychosis and externally validated their models on 79 patients with the same condition; however, this study was limited by the small sample size.

To overcome the two major limitations of these previous studies (no external validation and small sample size), we aimed to apply a machine learning approach using one of the largest longitudinal cohort studies of patients with first-episode psychosis (n=1027), for model development and internal validation, and data from a large randomised control trial (n=578) and two longitudinal cohort studies (totalling 162 patients), for external validation. We developed prediction models for multiple outcomes, including symptom remission and functional recovery (social recovery, vocational recovery, and quality of life) at 1 year after first-episode psychosis.

Methods

Study design and sources of data

In this machine learning approach, we used data from several sources: the National EDEN studies,[10] two Scottish validation datasets,[9,11] and the OPUS trial.[12] The National EDEN studies are a longitudinal naturalistic study of 1027 patients with first-episode psychosis recruited from 14 early intervention services across the National Health Service (NHS) in England (2005–10); the methods and baseline characteristics have been outlined previously.[10] The Scottish validation datasets were two longitudinal cohort studies of patients with first-episode psychosis: the Compassionate Recovery: Individualised Support in early Psychosis (CR:ISP) study[9] of 83 patients in NHS Greater Glasgow & Clyde (2011–14), and an earlier study[11] of 79 patients in NHS Glasgow and NHS Edinburgh (2006–09). The methodologies and baseline characteristics of these studies have been outlined previously.[9,11] The

OPUS trial[12] (NCT00157313) was a randomised controlled trial of 578 patients with first-episode psychosis recruited from all inpatient and outpatient mental health services in Copenhagen (Copenhagen Hospital Corporation) and Aarhus County, Denmark. OPUS assessed standard (n=272) versus specialised assertive intervention integrated treatment (n=275; January, 1998, to December, 2000). The methods and baseline characteristics of OPUS have been outlined previously.[12] Local ethics committees approved the studies and the trial. We have adhered to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement.[13]

With data from the EDEN studies, we developed four predictive models for each of the outcomes assessed in our study and internally validated the models by nested leave-one-site-out cross-validation (LOSOCV). Subsequently, we identified shared variables (from the top predictor variables from the internally validated models) between the EDEN studies and the Scottish datasets and between the EDEN studies and OPUS trial. We used these shared variables to build separate prediction models for external validation on the Scottish datasets and OPUS trial.

The key differences between EDEN (development) and Scottish (validation) datasets were the setting (EDEN was done in NHS England, whereas the Scottish datasets were from studies done in NHS Scotland) and study period (2005–10 in EDEN, 2011–14 and 2006–09 in the Scottish datasets). Both NHS England and NHS Scotland are free at the point of delivery. The key differences between EDEN and the other validation dataset (OPUS) were the setting (England *vs* Denmark, but both free at the point of delivery), study period (2005–10 *vs* January, 1998 to December, 2000 in OPUS), study type (naturalistic in EDEN, for which everyone received early intervention, *vs* randomised clinical trial in OPUS, for which early intervention was compared with treatment as usual), and inclusion criteria (participants were aged 14–35 years with a first presentation of psychotic symptoms in EDEN, whereas in OPUS, participants were aged 18–45 years with a diagnosis in the schizophrenia spectrum according to the International Classification of Diseases tenth edition codes in the F2 category, and participants in OPUS had not been given antipsychotic drugs for more than 12 weeks of continuous treatment). The inclusion and exclusion criteria for all three studies have been provided in the appendix (p 1).

Outcome variables

For EDEN, Scottish, and OPUS studies, assessments of predictors and outcomes were done by research assistants not directly involved in clinical care. We assessed four binary outcomes that were recorded at 1 year: symptom remission, meeting the Positive and Negative Syndrome Scale in Schizophrenia (PANSS) criteria at both 6 months and 1 year;[14] social recovery, achieving a Global Assessment of Functioning (GAF) score (range 0–100) of 65 or higher in EDEN, and a mean GAF symptoms and GAF disability score of 65 or higher in OPUS;[15] vocational recovery, assessing whether participants were in employment, education, or training;[16] and quality of life (QoL), assessed with the 3-level European QoL 5 Dimensions Index (EQ-5D-3L) time trade-off index based on UK population norms and dichotomised to greater than median (0·848)[17] in EDEN, the WHO QoL 26-item[18] instrument with total score dichotomised to greater than median (88) in the 2006–09 Scottish study, and the Lancashire QoL score[19] dichotomised at the median (43·5) in OPUS. We chose operationalised criteria for symptom remission[14] and included the three outcome measures for recovery to cover a broader patient-centred experience of recovery. Social recovery was not measured in the Scottish studies.

Statistical analysis

The EDEN study was powered for duration of untreated psychosis. OPUS was powered for positive symptoms

See Online for appendix

	Remission			Social recovery		
	Predictor (direction of effect)	In OPUS trial	In Scottish studies	Predictor (direction of effect)	In OPUS trial	In Scottish studies
1	PANSS P3—hallucinatory behaviour	✓	✓	GAF total	✓	
2	GAF total	✓		Main income source is salary or wage	✓	
3	Adjusted DUP in days	✓	✓	PAS client late adolescence social sexual aspects	✓	
4	Voluntary admission at baseline		✓	GAF disability total	✓	
5	PAS client late adolescence sociability withdrawal	✓	✓	Qualification level	✓	
6	PAS client general highest functioning achieved in life			History of ketamine use		
7	Qualification level	✓	✓	Main income source is state benefits	✓	
8	PAS client general energy level			PANSS P2—conceptual disorganisation	✓	
9	PANSS P2—conceptual disorganisation	✓	✓	PANSS P4—excitement		
10	Most serious self-harm is with premeditation of 3 h or less			History of amphetamine use	✓	
11	Hours a week doing leisure activities			PAS client general job change interrupted school attendance		
12	PANSS N4—passive social withdrawal	✓	✓	Atheist or agnostic	✓	
13	Housing type is own home or parents' home	✓	✓	PAS client late adolescence sociability withdrawal	✓	
14	Most serious self-harm is with knife or razor	✓		PANSS P3—hallucinatory behaviour	✓	
15	Community psychiatric nurse contact in last 3 months		✓	PAS client general employed or at school		
16	History of LSD use	✓		Voluntary admission at baseline		
17	History of ketamine use			PANSS P7—hostility	✓	
18	Any time spent per week doing leisure activities	✓		EQ-5D-3L health thermometer		
19	PAS client late adolescence social sexual aspects	✓	✓	PANSS N4—passive social withdrawal	✓	
20	PANSS G9—unusual thought content	✓	✓	Help by friends or relatives around the house in last 3 months	✓	
21	Was help sought in the prodromal phase?		✓	PANSS G11—poor attention	✓	
22	Insight scale awareness of symptoms	✓	✓	In paid employment at baseline		
23	Main income source is salary or wage	✓	✓	Community psychiatric nurse contact in last 3 months		
24	Help by friend or relative around the house in last 3 months	✓	✓	PANSS G8—uncooperativeness		
25	Any first degree relative with schizophrenia			Previous secondary psychiatric care		
26	Family member suggested care		✓	Mother tongue is language other than English but has good knowledge of English		
27	Years of schooling			GAF symptoms total	✓	
28	In education at baseline	✓	✓	EQ-5D-3L UK TTO index		
29	Most serious self-harm is with premeditation—not applicable			History of possible developmental disorder		
30	Never self-harmed	✓		Contact with criminal justice services in last 3 months	✓	
31				Any help from friends or relatives in last 3 months	✓	
32				Most serious self-harm is with overdose, drugs or alcohol		
33				First contact with EIS was facilitated by agency other than health, social care, criminal justice, or religious organisation		
34				PAS client general education		

(Figure 1 continues on next page)

	Vocational recovery Predictor (direction of effect)	In OPUS trial	In Scottish studies	Quality of life Predictor (direction of effect)	In OPUS trial	In Scottish studies
1	In employment, education, or training at baseline	✓	✓	EQ-5D-3L anxiety or depression		
2	PAS client general employed or at school			PAS client general job change interrupted school attendance		
3	Qualification level	✓	✓	PANSS G2—anxiety		✓
4	GAF disability total	✓	✓	History of amphetamine use	✓	
5	Main income source is salary or wage	✓	✓	EQ-5D-3L mobility		
6	In education at baseline	✓	✓	PANSS P3—hallucinatory behaviour	✓	✓
7	Main income source is state benefits	✓	✓	PAS client early adolescence sociability withdrawal	✓	✓
8	Length of time since most recent self-harm			Housing type is own home or parents' home	✓	✓
9	Any time spent per week doing childcare activities		✓	PAS client early adolescence social sexual aspects	✓	✓
10	PAS client general degree of interest in life			EQ-5D-3L health thermometer		
11	PAS client general education			Education level	✓	✓
12	Any time spent per week doing sport activities	✓		Qualification level	✓	
13	In voluntary employment at baseline			Any first degree relative with schizophrenia		
14	Ethnicity—white British	✓	✓	First contact with EIS was with police		✓
15	First contact with EIS was with police		✓	Main income source is salary or wage	✓	✓
16	In paid employment at baseline			PANSS G11—poor attention	✓	✓
17	Calgary Depression Scale total	✓	✓	EQ-5D-3L UK TTO index		
18	GAF total	✓		PAS client general energy level		
19	PANSS P2—conceptual disorganisation	✓	✓	PANSS P2—conceptual disorganisation	✓	✓
20	Main income source is something other than family, salary, or benefits			Initial appointment attended by client and family		
21	PANSS G13—disturbance of volition	✓	✓	PANSS P5—grandiosity	✓	✓
22	Most serious self-harm is with knife or razor	✓		PANSS G6—depression	✓	✓
23	Adjusted DUP in days	✓	✓	GAF total	✓	
24	PAS client childhood sociability withdrawal	✓	✓	History of ketamine use		
25	GAF symptoms total	✓		History of cocaine use	✓	✓
26	Ethnicity—Pakistani		✓	Number of previous admissions		
27	Housing type is own home or parents' home	✓	✓	PAS client late adolescence sociability withdrawal	✓	✓
28	Sleeps for 8 h or more each day			Main income source is state benefits	✓	✓
29	Family member suggested care		✓	Any time spent per week doing housework activities		
30	PAS client general job change or interrupted school attendance			PANSS P6—suspiciousness		✓
31	In receipt of any state benefits			Living with parents or guardian	✓	✓
32	PANSS N6—lack of spontaneity	✓	✓	Most serious self-harm is with knife or razor	✓	
33	Help by friend or relative around the house in last 3 months	✓	✓	Most serious self-harm is with overdose, drugs, or alcohol		
34	Housing type is rented		✓	PANSS P7—hostility	✓	✓
35	Most serious violence victim gender was male			Never self-harmed	✓	
36	Any help by friend or relative in last 3 months	✓	✓	PANSS N6—lack of spontaneity	✓	✓
37	Contact with criminal justice services in last 3 months	✓		Help by friend or relative around the house in last 3 months	✓	
38	PANSS G11—poor attention	✓	✓	Housing type is rented		✓
39				Adjusted DUP in days	✓	✓
40				PANSS N5—difficulties in abstract thinking		✓
41				Male sex	✓	✓
42				History of LSD use	✓	
43				Number of second-degree relatives with a psychiatric family history		
44				History of cannabis use	✓	✓

Figure 1: **Top prediction variables for each outcome**
Top predictor variables selected by elastic net regularisation across all 14 LOSOCV models for each outcome, ordered by their mean rank across the 14 models by absolute coefficient magnitude, along with their direction of effect (red is negative, grey is positive). PANSS=Positive And Negative Symptom Scale. GAF=Global Assessment of Functioning scale. DUP=Duration of Untreated Psych====osis. PAS=Premorbid Adjustment Scale. LSD=Lysergic acid diethylamide. EQ-5D-3L=3-level European Quality of Life 5 Dimensions Index. UK TTO=time trade-off index based on UK population norms. EIS=Early Intervention Service. LOSOCV=leave-one-site-out cross-validation.

according to the Scale for Assessment of Positive Symptoms (SAPS). The 2006–09 Scottish study was powered for the strength of association between duration of untreated psychosis and psychiatric symptomatology. The 2011–14 Scottish study was powered for positive and negative symptoms. Because our study is a post-hoc analysis, a sample size calculation is not applicable.

Studies with missing outcome data were removed from the analysis. Regarding predictor selection, during data pre-processing in EDEN, all 266 baseline social, demographic, and clinical predictor variables were centred and scaled, variables with zero variance and near-zero variance were removed, and variables with more than 20% of missing data were excluded. For the remaining 163 (61%) of predictor variables (appendix pp 2–6), missing data were imputed by use of *k*-nearest neighbour imputation (k=5) to increase prediction performance.[20] We did not complete any a-priori hypothesis-based feature selection.

We used the EDEN dataset for model development and undertook both internal and external validation with LOSOCV.[21] We fit a logistic regression model by elastic net regularisation with variable selection in the caret package[22] using the glmnet package.[23] Glmnet fits a generalised linear model through penalised maximum likelihood (appendix pp 1–2). All the 163 predictor variables were used simultaneously with the elastic net regularisation model. Each of the 14 EDEN sites was left out once for the validation of a model based on the remaining 13 sites and trained by use of a ten-fold cross-validation (splits balanced by outcome class) over a 10×10 grid of α and λ hyperparameters, with Breiman's 1 SE rule.[24]

We measured average performance across the resulting 14 best LOSOCV models using receiver operating characteristic (ROC) curve and area under the curve (AUC). AUCs, with 95% CIs, were established on the basis of U-statistic theory, and permutation testing confirmed significance. Representative model accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), prognostic summary index (PSI), positive likelihood ratio (LR+) and negative likelihood ratio (LR–) are presented on the basis of the point on the ROC curve corresponding to Youden's index. We assessed the stability (ϕ) of feature selection in the 14 best LOSOCV models using the approach described by Nogueira and colleagues,[25] where ϕ lower than 0·4 shows poor agreement between the 14 models, 0·4 to 0·75 shows intermediate to good agreement, and higher than 0·75 shows excellent agreement. We did this model development procedure for each of our four binary outcomes.

We assessed the relatedness of the four models by computing the Yule φ correlation between the four outcomes, computing the Pearson correlation between probability outputs of the four logistic regression models, and assessing the prediction performance when using the probability outputs of one model as predictors of outcome for the other three models with LOSOCV (appendix pp 6–7). We used the shared predictor variables among the top variables for the four models to build generalised linear models for external validation.

For external (geographical and temporal) validation of the prediction models, we used the Scottish and OPUS datasets. For each outcome, we took the shared variables across both the EDEN and the external validation dataset from the top predictor variables determined during model development (those selected in all 14 LOSOCV models; figure 1). We standardised these variables

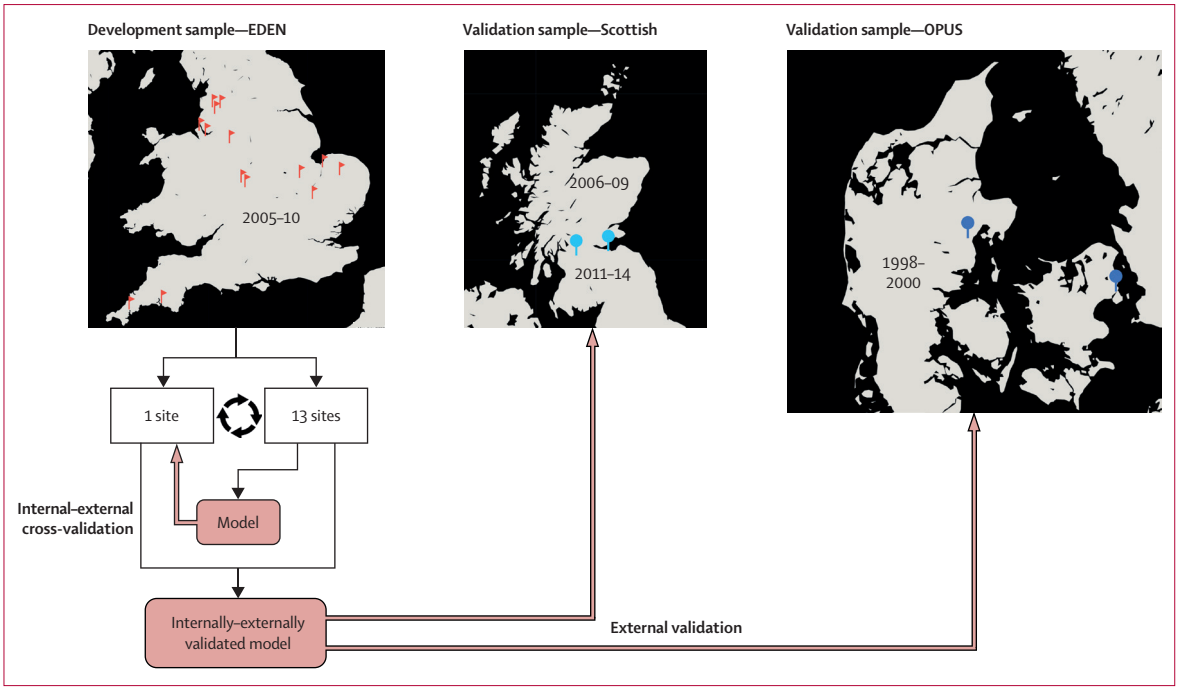

Figure 2: **Analysis pipeline**
Elastic net model development and internal–external validation using a leave-one-site-out cross-validation in the EDEN sample. Internally–externally validated generalised linear models were constructed with use of top predictors shared between the EDEN and Scottish datasets, and the EDEN and OPUS datasets. These were then externally validated on the Scottish datasets and the OPUS dataset.

separately on each dataset before model fitting; therefore, we were able to assess EDEN model performance on the validation dataset even though some shared variables were measured on different scales. Afterwards, we used the entire EDEN dataset to fit a generalised linear model by maximum likelihood estimation (without regularisation) using these shared top predictor variables (having found no improvement in performance during initial scoping with more complex classifiers, including linear and radial support vector machines, elastic net, and random forest). We confirmed that the internal–external validation performance on the EDEN dataset remained robust with the new model using only the shared top predictor variables. The internally–externally validated EDEN model was then externally validated on the external dataset, with performance reported as already outlined. This process was repeated separately for the Scottish datasets and the OPUS dataset (figure 2).

All statistical analyses were done with R, and the code is available online. The comparison between EDEN, Scottish, and OPUS samples (demographic and social variables) is provided in the appendix (p 2).

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

	Training data (EDEN studies)	Validation data (Scottish studies)	Validation data (OPUS trial)	p value
Symptom remission	320/673 (48%)	66/131 (50%)	121/338 (36%)	0·0006*
Social recovery	388/829 (47%)	NA	73/518 (14%)	<0·0001*
Vocational recovery	436/807 (54%)	59/142 (42%)	173/553 (31%)	<0·0001*
Quality of life	328/729 (45%)	23/47 (49%)	113/226 (50%)	0·39

Data are n/N (%). NA=not applicable. *Significant differences (determined with Pearson's χ^2 test) of rates of positive outcomes between the cohorts, after Holm-Bonferroni correction.

Table 1: **Outcome data for training and validation cohorts**

Results

We included only participants for whom outcome data were available (table 1). In the EDEN studies, 673 (66%) of 1027 patients had complete symptom remission outcome data, 829 (81%) had complete social recovery outcome data, 807 (79%) had complete vocational recovery outcome data, and 729 (71%) had complete QoL outcome data. In the Scottish studies, 131 (81%) of 162 patients had complete symptom remission outcome data, 142 (88%) had complete vocational recovery outcome data, and 47 (59%) of 79 had complete QoL outcome data. In the OPUS trial, 338 (58%) of 578 patients had complete symptom remission outcome data, 518 (90%) had complete social recovery outcome data, 553 (96%) had complete vocational recovery outcome data, and 226 (39%) had complete QoL outcome data. 15–39% of patients were missing outcomes data on model performance at 1 year for training cohorts

For the **R code** see https://github.com/samleighton87/EDEN_R_Code

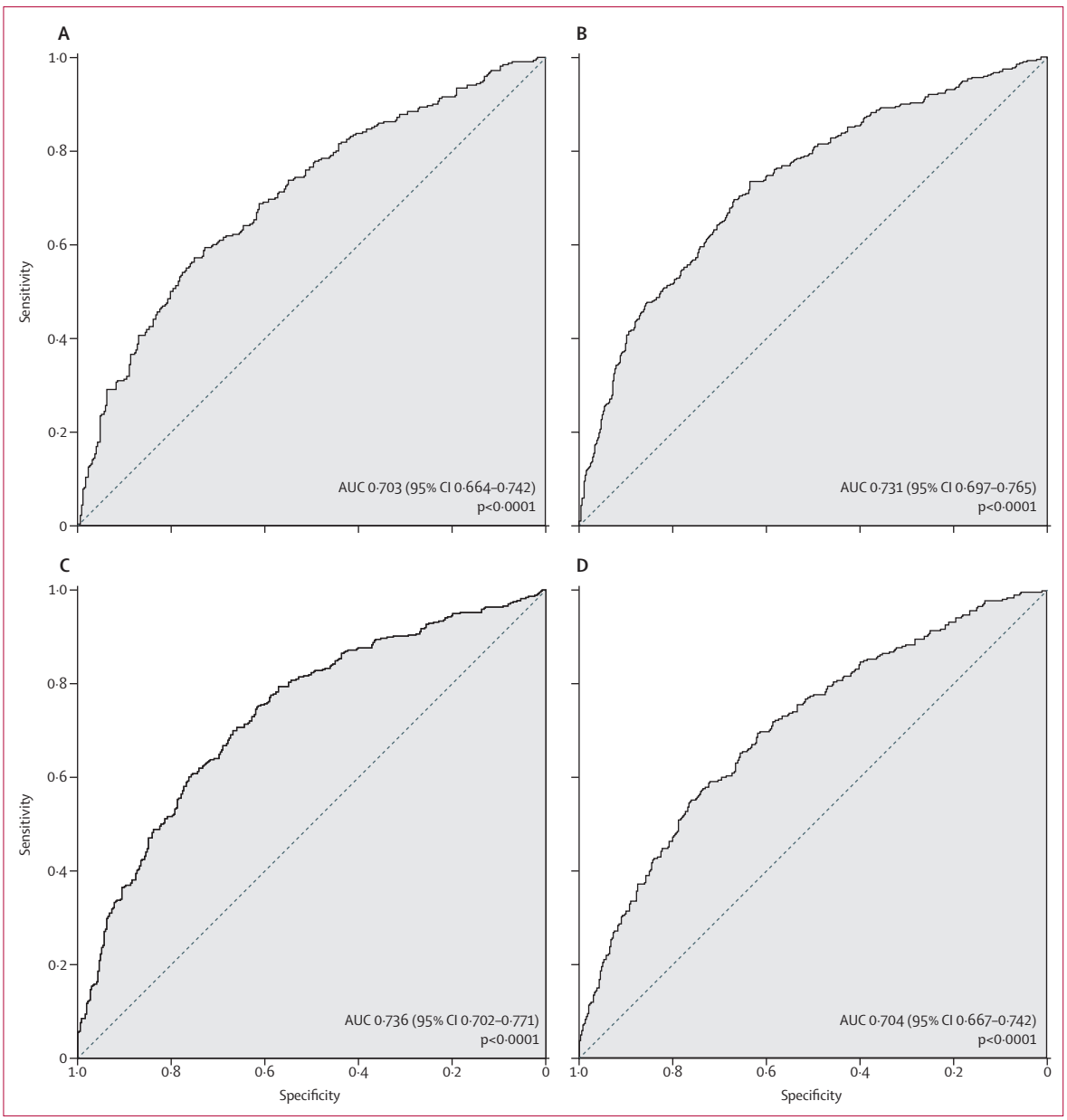

Figure 3: **ROC curves showing internal–external LOSOCV model performance in the EDEN dataset for 1-year symptom recovery (A), social recovery (B), vocational recovery (C), and quality of life (D) models**
ROC=receiver operating characteristic. LOSOCV=leave-one-site-out cross-validation. AUC=area under the curve.

and for validation cohorts these values were 4–61% for OPUS and 19–61% for Scottish studies.

During internal cross-validation with all the 163 predictor variables simultaneously, all of our four models had AUCs higher than 0·700, significantly better than chance (p<0·0001; figure 3, table 2). The accuracy achieved by the four models was higher than AUC 0·65, and the PSI of the four models was higher than 0·31, indicating a 31% additional gain in prediction certainty.[26] The stability of feature selection in the 14 LOSOCV models was 0·54 for the remission model, 0·67 for the social recovery model, 0·71 for the vocational recovery model, and 0·70 for the QoL model (appendix p 7).

The correlation of the probability outputs of the four models was higher than the correlation of the respective outcomes that they were trained to predict. Each model predicted its outcome best, but they also significantly predicted each of the other three outcomes, with a lower level of performance (appendix pp 6–7).

The top predictors for the four models selected by the elastic net model are provided in figure 1. The four models included predictor variables ranging from demographic

characteristics, family history, premorbid functioning, baseline education and employment status, social factors, duration of untreated psychosis, and baseline symptoms (appendix pp 2–6). These models had similar performance to the elastic net model built with use of the 163 predictor variables on LOSOCV. The external validation performance of the generalised linear models was significantly better than chance, with AUCs higher than 0·67 for symptom remission and QoL outcomes, and higher than 0·86 for the vocational recovery outcome in the Scottish datasets. The external validation performance of the generalised linear models had AUCs of 0·61 for remission, 0·57 for social recovery, and 0·66 for vocational recovery outcomes in the OPUS dataset. The AUC of the generalised linear model for QoL was not statistically significant (table 2, appendix pp 11–12). We did external validation performance for the two groups of the OPUS trial (appendix pp 9–10). Model performance was better in the standard treatment group for remission and social recovery than for the other outcomes, whereas in the intervention group, performance was better for vocational recovery.

Discussion

In this study, we developed outcome prediction models for remission and recovery for people with first-episode psychosis using baseline sociodemographic and clinical variables, and we internally cross-validated the models with a large naturalistic cohort study (EDEN study). We externally validated the prediction models on patients from three studies: two longitudinal cohort studies of patients with first-episode psychosis (Scottish studies) and a randomised control trial of specialised assertive intervention treatment versus standard treatment (OPUS). The predictive performance of the models were in the range of values for established calculators in use for predicting risk of cardiovascular diseases (AUC 0·71–0·76)[27] and cancer (0·57–0·72).[28–30] The PSIs indicated that our prediction models provided a 31–37% increase in prognostic certainty compared with that of pre-test probabilities at 1 year.

We developed prediction models for multiple outcomes, including remission and recovery (social and vocational recovery and QoL), in recognition of the fact that intervention strategies might be distinct for each outcome,[31] even though each of our models was able to accurately predict other outcomes significantly better than chance, albeit with reduced performance. Our prediction model for social recovery had similar performance (AUC 0·731) to that of Koutsouleris and colleagues'[8] model (balanced accuracy 0·71), though their study was limited by the absence of a true external validation. Our model performance for the remission outcome (AUC 0·703 [95% CI 0·664–0·742]) was better than that of Leighton and colleagues'[9] model (0·635–0·670), whereas their model performance for vocational recovery was better than that of our model. In our study, the stability of feature

	Symptom remission		Social recovery		Vocational recovery			Quality of life	
	Internal–external validation (EDEN)	External validation (Scottish)	External validation (OPUS)	Internal–external validation (EDEN)	External validation (OPUS)	Internal–external validation (EDEN)	External validation (Scottish)	External validation (OPUS)	Internal–external validation (EDEN)	External validation (Scottish)	External validation (OPUS)
Performance	0·703 (0·664–0·742; p<0·0001)	0·680 (0·587–0·773; p=0·0004)	0·616 (0·553–0·679; p=0·0003)	0·731 (0·697–0·765; p<0·0001)	0·573 (0·504–0·643; p=0·04)	0·736 (0·702–0·771; p<0·0001)	0·867 (0·805–0·930; p<0·0001)	0·660 (0·610–0·710; p<0·0001)	0·704 (0·667–0·742; p<0·0001)	0·679 (0·522–0·836; p=0·03)	0·556 (0·481–0·631; p=0·07)
Accuracy	0·670 (0·636–0·703)	0·695 (0·618–0·771)	0·618 (0·524–0·704)	0·687 (0·657–0·718)	0·456 (0·328–0·817)	0·693 (0·660–0·725)	0·838 (0·775–0·894)	0·680 (0·609–0·725)	0·668 (0·632–0·704)	0·702 (0·596–0·809)	0·589 (0·540–0·637)
Sensitivity	0·584 (0·491–0·827)	0·621 (0·455–0·773)	0·612 (0·306–0·843)	0·722 (0·487–0·778)	0·781 (0·233–0·945)	0·722 (0·573–0·821)	0·898 (0·780–0·966)	0·584 (0·457–0·723)	0·623 (0·512–0·774)	0·957 (0·564–1·000)	0·876 (0·419–0·947)
Specificity	0·751 (0·544–0·827)	0·769 (0·615–0·908)	0·629 (0·378–0·885)	0·660 (0·594–0·871)	0·396 (0·234–0·906)	0·666 (0·550–0·803)	0·807 (0·699–0·904)	0·726 (0·574–0·824)	0·711 (0·551–0·803)	0·500 (0·250–0·833)	0·301 (0·204–0·743)
PPV	0·679 (0·601–0·739)	0·729 (0·636–0·854)	0·476 (0·412–0·636)	0·650 (0·616–0·769)	0·179 (0·158–0·333)	0·719 (0·673–0·785)	0·766 (0·679–0·867)	0·490 (0·421–0·563)	0·633 (0·575–0·701)	0·640 (0·561–0·800)	0·559 (0·527–0·642)
NPV	0·667 (0·631–0·734)	0·667 (0·593–0·759)	0·742 (0·687–0·829)	0·726 (0·655–0·766)	0·914 (0·876–0·967)	0·668 (0·606–0·736)	0·911 (0·840–0·971)	0·793 (0·760–0·831)	0·700 (0·659–0·759)	0·900 (0·643–1·000)	0·706 (0·555–0·841)
PSI	0·346 (0·232–0·473)	0·396 (0·229–0·613)	0·217 (0·099–0·465)	0·376 (0·271–0·535)	0·093 (0·034–0·300)	0·387 (0·279–0·521)	0·677 (0·519–0·838)	0·283 (0·181–0·394)	0·333 (0·234–0·460)	0·540 (0·204–0·800)	0·265 (0·081–0·483)
LR+	2·345 (1·077–4·497)	2·688 (1·182–8·402)	1·649 (0·491–7·318)	2·124 (1·120–6·031)	1·292 (0·304–10·013)	2·162 (1·273–4·168)	4·653 (2·591–10·063)	2·133 (1·071–4·098)	2·156 (1·140–3·929)	1·914 (0·752–5·988)	1·253 (0·526–3·690)
LR–	0·554 (0·268–0·936)	0·493 (0·250–0·886)	0·617 (0·177–1·838)	0·421 (0·255–0·864)	0·554 (0·061–3·282)	0·417 (0·223–0·776)	0·126 (0·038–0·315)	0·573 (0·337–0·947)	0·530 (0·281–0·886)	0·086 (0·000–1·744)	0·412 (0·071–2·856)

Data are AUC (95% CI) or AUC (95% CI; p value). Leave-one-site-out internal-externally validated performance and external validation metrics for year 1 for binary symptomatic, social recovery, vocational recovery and internal-external validated performance for quality of life outcome metrics. 95% CIs of the AUC were established on the basis of U-statistic theory, and significance level confirmed by permutation testing (n=10001). Representative accuracy, sensitivity, specificity, PPV, NPV, PSI, LR+, and LR– with bootstrapped 95% CIs (n=2000) are based on the point on the receiver operating characteristic curve corresponding to Youden's index. Different thresholds can be taken on the basis of the requirements of the diagnostic test. AUC=area under the curve. PPV=positive predictive value. NPV=negative predictive value. PSI=prognostic summary index. LR+=positive likelihood ratio. LR–=negative likelihood ratio.

Table 2: **Performance metrics for internal–external and external validation**

selection in the 14 LOSOCV models (0·54 for the remission model, 0·67 for the social recovery model, 0·71 for the vocational recovery model, and 0·70 for the QoL model) indicates an intermediate to good strength of agreement within each of the four prediction models.[25]

The external validation performance of the prediction models was similar to that of the training dataset for the Scottish datasets, although the performance was reduced in the OPUS dataset. Several possible explanations exist for this difference. The external validation models were necessarily built with use of shared variables alone, not with all the top identified predictor variables. However, a repeat internal validated LOSOCV performance with models using just the shared variables remained similar in the EDEN dataset. The way outcomes (and some predictors) were measured differed between the datasets: remission was defined with Andreasen criteria, but with use of PANSS for EDEN and Scottish datasets and SAPS–SANS (Scale for the Assessment of Negative Symptoms) for the OPUS dataset; social functioning was defined by GAF for EDEN, but by use of the mean of GAF symptoms and GAF disability for OPUS; and for QoL, EDEN used EQ-5D-3L, but the other three datasets used WHO QoL. The measurement of vocational recovery outcomes was similar across the datasets, and the fact that this model performed best in external validation could reflect this. Furthermore, we found significant differences in the balance of remission and recovery outcome rates between all datasets, although the OPUS dataset had much fewer remission and recovery outcomes than those of the other datasets. This finding might be explained by the differing timeframes of data collection and the fact that the EDEN and Scottish datasets were collected from patients in early-intervention services, whereas OPUS was a randomised controlled trial of intensive versus standard treatment. Contrary to our expectation, the validation performance was better for the remission and social recovery outcomes for the standard treatment group of the OPUS trial. The validation performance for the vocational recovery model was better for the intensive treatment group, which is similar to the performance in the training dataset. Taken together, these issues are unavoidable in the context of our analyses being opportunistic and post-hoc, with use of existing datasets. However, the fact that model performance was significantly better than chance on external validation (except for QoL in OPUS), despite these differences, is very promising for the ability of such methods to withstand heterogeneous data in real-world clinical settings.

Our analysis has several strengths. The data for the model development were derived from one of the largest naturalistic cohort studies in patients with first-episode psychosis treated in early-intervention services. We used LOSOCV for model development and internal validation. We found the stability of the feature selection with LOSOCV for 14 sites to have intermediate to good level of agreement. Furthermore, we externally validated the models in three independent datasets with different time periods, geographical regions, and recording methods. We used strict operationalised outcome criteria to define symptomatic outcomes and developed prediction models for multiple outcomes. Each of the individual prediction models predicted the other three outcomes better than chance, although with reduced performance. An argument exists for using one prediction model to predict multiple outcomes, although this would come with a trade-off of marginally reduced performance and needs further testing in prospective clinical trials.

Our study also has several limitations. About 49% of eligible patients consented to participate in the EDEN study, which might affect the generalisability of our prediction models to all patients with first-episode psychosis. However, participants who did not consent had characteristics largely similar at baseline to those of individuals who consented to participate.[10] Despite this, we cannot assume that the models developed with data from the patients included in the EDEN study would have a better performance than chance in individuals not included in the EDEN study sample. The effect of missing outcomes data on model performance was not trivial in patients at 1 year for training cohorts (15–39%) and for validation cohorts (OPUS 4–61%; Scottish studies 19–61%). This effect might introduce bias and affect the generalisability of our results. Importantly, our models have not been validated for prediction after baseline as treatment progresses. Future studies could consider building models that account for change over time or in response to treatment (eg, dynamic Bayesian networks with continuous retraining). We did not collect cognitive and physical biomarkers of illness, including blood samples and neuroimaging, which previous studies have highlighted as potentially important for generating accurate predictions.[32] The duration criteria for recovery has been proposed to be at least 2 years.[33] However, the criteria used in our analysis for recovery outcomes were much narrower and, for three of four measures, were based on point outcomes (GAF; employment, education, or training status; and QoL) at 1 year. The prevalence of recovery in our training cohort was similar to that reported in a large meta-analysis,[4] but higher than that reported in another meta-analysis,[2] which might also affect the generalisability of our model.

The decision making process to determine which interventions to use and for how long in the treatment of patients with a first-episode psychosis is based on clinical intuition. We are not aware of evidence assessing how accurate clinicians using baseline information are at predicting 1-year outcomes for psychosis, although it has been shown that clinicians are poor at predicting outcomes in depression.[34,35] Clinicians working with patients with first-episode psychosis might benefit from a reliable and methodologically robust tool to identify individuals with likelihood of a good or poor outcome at initial clinical contact, so that the information on

outcome prediction can be used alongside clinical judgment for stratification of treatment.

Patients with good outcomes are likely to need a different set of interventions and duration of treatment compared with patients with poorer outcomes. If outcome prediction models are developed into clinically applicable tools after further rigorous testing of their usefulness in a prospective clinical trial, they could assist in clinical decision making, leading to better use of clinical resources by providing targeted interventions based on individual predictions of patient outcomes. Guidelines could be developed in consultation with stakeholders on how to put such tools into practice to facilitate a stepped model of care. Future work should identify, in a prospective clinical trial, whether one prediction model might accurately predict multiple outcomes and whether it is possible to update the prediction models prospectively over time and in response to different interventions. Whether the addition of other predictors, including biomarkers, will improve prediction accuracy of the models remains to be tested.

In our machine learning analysis of a longitudinal cohort of patients with first-episode psychosis treated in early-intervention services, we were able to show that multiple outcomes can be reliably predicted for patients by use of baseline demographic and clinical variables at 1 year, with external generalisability. Our prediction models have similar discriminatory power to other available predictive models.[8,9] Our models benefit from being developed with use of a naturalistic cohort study and externally validated in a cohort study and a randomised control trial, together with the use of readily available clinical data and, to our knowledge, the largest sample size used in a machine learning study of first-episode psychosis to date. Furthermore, to our knowledge, our study represents the first published evidence for the use of machine learning models of QoL outcome in patients with a first-episode psychosis.

Contributors

PKM conceptualised the analysis plan. PKM and SPL designed the analysis plan and drafted and revised the paper. SPL did the analysis. RU, MRB, RK, JC, GVG, MS, and PFL contributed to the interpretation of the analysis. RU, MRB, RK, JC, GVG, MS, PFL, SPS, LE, PBJ, DF, VS, NF, AIG, MEB, and RHBC revised the draft. SPS, LE, PBJ, DF, VS, and NF designed the EDEN study. SPS, LE, PBJ, DF, VS, NF, AIG, MS, MEB, and RHBC contributed data. MEB, RHBC, and NF did the validation analysis on OPUS data. MB monitored data collection for the EDEN studies and supervised them. AIG monitored the data collection for the Scottish studies and supervised them. MN monitored the data collection for the OPUS trial and supervised it. MB, AIG, and MN revised and approved the final version of the manuscript.

Declaration of interests

PKM has received honorariums from Sunovion and Sage. RU has received honorariums from Sunovion. JC has received grants from Wellcome Trust and Sackler Trust and honorariums from Johnson & Johnson. SPS and MB are part-funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care WM (CLAHRC-WM). GVG has received support from H2020-EINFRA, the NIHR Birmingham ECMC, NIHR Birmingham SRMRC, the NIHR Birmingham Biomedical Research Centre, and the MRC HDR UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England), the devolved administrations, and leading medical research charities. All other authors declare no competing interests.

Data sharing

MB acts as custodian of the EDEN dataset and data sharing and secondary analyses are supported under the auspices of the University of Warwick (Coventry, UK); please contact MB for all requests. AIG acts as custodian of the Scottish datasets and data sharing and secondary analyses are supported under the auspices of the University of Glasgow (Glasgow, UK); please contact AIG for all requests. MN acts as custodian of the OPUS trial dataset and data sharing and secondary analyses are supported under the auspices of the University of Copenhagen (Copenhagen, Denmark); please contact MN for all requests.

Acknowledgments

EDEN was funded by the UK Department of Health (RDD/ARF2) and National Institute of Health Research under the Programme Grants for Applied Research Programme (RP-PG-0109–10074). The Scottish studies were funded by the National Health Service Research Scotland (NRS), through the Chief Scientist Office (CZH/4/295 and CZH/3/5), the Scottish Mental Health Research Network, and the Wellcome Trust (104025/Z/14/Z). The OPUS trial was funded by the Danish Ministry of Health (jr.nr. 96-0770-71), Danish Ministry of Social Affairs, University of Copenhagen, Copenhagen Hospital Corporation, Danish Medical Research Council (jr.nr. 9601612 and 9900734), and Slagtermester Wørzners Fond. This study and external validation were part funded by an unrestricted grant from The Lundbeck Foundation to PRECISE (R277–2018–1411). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR, the Medical Research Council, the UK Department of Health, or the CLAHRC-WM collaborative organisations. We thank all participants of the EDEN, OPUS, and Scottish studies and the EIP teams who supported this research. We thank the anonymous reviewers for providing very helpful reviews to improve the manuscript.

References

1 Kirkbride JB, Errazuriz A, Croudace TJ, et al. Incidence of schizophrenia and other psychoses in England, 1950–2009: a systematic review and meta-analyses. *PLoS One* 2012; **7:** e31660.

2 Lally J, Ajnakina O, Stubbs B, et al. Remission and recovery from first-episode psychosis in adults: systematic review and meta-analysis of long-term outcome studies. *Br J Psychiatry J Ment Sci* 2017; **211:** 350–58.

3 Marwaha S, Thompson A, Upthegrove R, Broome MR. Fifteen years on—early intervention for a new generation. *Br J Psychiatry J Ment Sci* 2016; **209:** 186–88.

4 Jääskeläinen E, Juola P, Hirvonen N, et al. A systematic review and meta-analysis of recovery in schizophrenia. *Schizophr Bull* 2013; **39:** 1296–306.

5 Santesteban-Echarri O, Paino M, Rice S, et al. Predictors of functional recovery in first-episode psychosis: a systematic review and meta-analysis of longitudinal studies. *Clin Psychol Rev* 2017; **58:** 59–75.

6 Malla A, Payne J. First-episode psychosis: psychopathology, quality of life, and functional outcome. *Schizophr Bull* 2005; **31:** 650–71.

7 Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 2016; **375:** 1216–19.

8 Koutsouleris N, Kahn RS, Chekroud AM, et al. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *Lancet Psychiatry* 2016; **3:** 935–46.

9 Leighton SP, Krishnadas R, Chung K, et al. Predicting one-year outcome in first episode psychosis using machine learning. *PLoS One* 2019; **14:** e0212846.

10 Birchwood M, Lester H, McCarthy L, et al. The UK national evaluation of the development and impact of Early Intervention Services (the National EDEN studies): study rationale, design and baseline characteristics. *Early Interv Psychiatry* 2014; **8:** 59–67.

11 Gumley AI, Schwannauer M, Macbeth A, et al. Insight, duration of untreated psychosis and attachment in first-episode psychosis: prospective study of psychiatric recovery over 12-month follow-up. *Br J Psychiatry J Ment Sci* 2014; **205:** 60–67.

12 Petersen L, Nordentoft M, Jeppesen P, et al. Improving 1-year outcome in first-episode psychosis: OPUS trial. *Br J Psychiatry* 2005; **187:** s98–103.

13 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; **350:** g7594.

14 Andreasen NC, Carpenter WT, Kane JM, Lasser RA, Marder SR, Weinberger DR. Remission in schizophrenia: proposed criteria and rationale for consensus. *Am J Psychiatry* 2005; **162:** 441–49.

15 Endicott J, Spitzer RL, Fleiss JL, Cohen J. The Global Assessment Scale: a procedure for measuring overall severity of psychiatric disturbance. *Arch Gen Psychiatry* 1976; **33:** 766–71.

16 International First Episode Vocational Recovery (iFEVR) Group. Meaningful lives: supporting young people with psychosis in education, training and employment: an international consensus statement. *Early Interv Psychiatry* 2010; **4:** 323–26.

17 Kind P, Hardman G, Macran S. UK population norms for EQ-5D. 1999. https://EconPapers.repec.org/RePEc:chy:respap:172chedp (accessed July 16, 2018)

18 Skevington SM, Lotfy M, O'Connell KA, WHOQOL Group. The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. A report from the WHOQOL group. *Qual Life Res* 2004; **13:** 299–310.

19 van Nieuwenhuizen Ch, Schene AH, Koeter MWJ, Huxley PJ. The Lancashire Quality of Life Profile: modification and psychometric evaluation. *Soc Psychiatry Psychiatr Epidemiol* 2001; **36:** 36–44.

20 Liu Y, Gopalakrishnan V. An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. *Data* 2017; **2:** 8.

21 Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016; **69:** 245–47.

22 Kuhn M. Building predictive models in R using the caret package. 2008. https://www.jstatsoft.org/article/view/v028i05 (accessed July 18, 2018)

23 Friedman J, Hastie T, Tibshirani R. Regularization paths for Generalized linear models via coordinate descent. *J Stat Softw* 2010; **33:** 1–22.

24 Breiman L. Classification and regression trees. New York: Routledge, 1984.

25 Nogueira S, Sechidis K, Brown G. On the stability of feature selection algorithms. *J Mach Learn Res* 2018; **18:** 1–54.

26 Linn S, Grunau PD. New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests. *Epidemiol Perspect Innov* 2006; **3:** 11.

27 Bitton A, Gaziano TA. The Framingham Heart Study's impact on global risk assessment. *Prog Cardiovasc Dis* 2010; **53:** 68–78.

28 Pfeiffer RM, Park Y, Kreimer AR, et al. Risk prediction for breast, endometrial, and ovarian cancer in white women aged 50 y or older: derivation and validation from population-based cohort studies. *PLoS Med* 2013; **10:** e1001492.

29 Specht MC, Kattan MW, Gonen M, Fey J, Van Zee KJ. Predicting nonsentinel node status after positive sentinel lymph biopsy for breast cancer: clinicians versus nomogram. *Ann Surg Oncol* 2005; **12:** 654–59.

30 Kattan MW, Yu C, Stephenson AJ, Sartor O, Tombal B. Clinicians versus nomogram: predicting future technetium-99m bone scan positivity in patients with rising prostate-specific antigen after radical prostatectomy for prostate cancer. *Urology* 2013; **81:** 956–61.

31 Cannon TD, Yu C, Addington J, et al. An individualized risk calculator for research in prodromal psychosis. *Am J Psychiatry* 2016; **173:** 980–88.

32 Upthegrove R, Manzanares-Teson N, Barnes NM. Cytokine function in medication-naive first episode psychosis: a systematic review and meta-analysis. *Schizophr Res* 2014; **155:** 101–08.

33 Faerden A, Nesvåg R, Marder SR. Definitions of the term 'recovered' in schizophrenia and other disorders. *Psychopathology* 2008; **41:** 271–78.

34 Leuchter AF, Cook IA, Marangell LB, et al. Comparative effectiveness of biomarkers and clinical indicators for predicting outcomes of SSRI treatment in major depressive disorder: results of the BRITE-MD study. *Psychiatry Res* 2009; **169:** 124–31.

35 Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* 2016; **3:** 243–50.